The role of known genetic variants in the base quality score recalibration for the genetic variant calling

Sunhee Kim, Dongju Lee and Chang-Yong Lee*

Department of Industrial and Systems Engineering, Kongju National University

*Corresponding author: clee@kongju.ac.kr

ABSTRACT

The genetic variant calling from genome data relies on various pipelines of computational tools to account for systematic differences in the genome data of different species. In particular, the base quality score recalibration (BQSR) in the pipeline is a pre-processing step that leverages a large database of known variants called dbSNP. While these pipelines are expected to be applicable in a species-independent manner, they have not been carefully evaluated with non-human data. To investigate the impact of the dbSNP on BQSR, we analyzed genomic sequencing data from four different species: human, sheep, rice, and chickpea. Unexpectedly, the recalibrated scores and the error rate obtained by BQSR were biased by the size of the dbSNP and its builds. To address this issue, we suggest an alternative to the dbSNP by constructing a pseudo-database for various species based on the sequence data.

I. Introduction

- Quality of a base can be expressed in terms of Phred score as

$$Q_{Phred} = -10\log_{10} p_\varepsilon$$

where p_ε is the error rate. Q_{Phred} is an integer and known as the base quality score.

- Because the correct estimate of the quality score is essential in the variant calling, GATK provides a recalibration tool for the base quality score (BQSR).
- BQSR is a method of adjusting platform-provided base quality scores to be more accurate by using an external resource of known variants, such as the dbSNP.
- The usefulness of BQSR heavily relies on the number and quality of reported variants in the database. So, when the database is incomplete, mismatched bases are less likely to be identified by database.
- We proposed a recalibration method when the number of variants in a databases is not enough, and compared the result of the dbSNP with the proposed method.

II. Materials and method

1. Data acquisition

Human: FASTQ : 1000 Genomes Project

- reference sequence : GRCh38(Genome Reference Consortium Human Build 38)
- dbSNP : dbSNP Build 151(Homo sapiens) of 634M variants

Rice: FASTQ : 3000 Rice Genomes Project

- reference sequence : Nipponbare IRGSP-1.0
- dbSNP : dbSNP Build 151(Oryza sativa) of 12M variants

Sheep: FASTQ : International Sheep Genomics Consortium

- reference sequence : Oar_rambouillet_v1.0
- dbSNP : dbSNP Build 151(Ovis aries) of 68M variants

Chickpea: FASTQ : International Crops Research Institute for the Semi-Arid Tropics

- reference sequence : Car_ref_ASM33114V1
- dbSNP : dbSNP Build 146(Cicer arietinum) of 327K variants

2. Data analysis

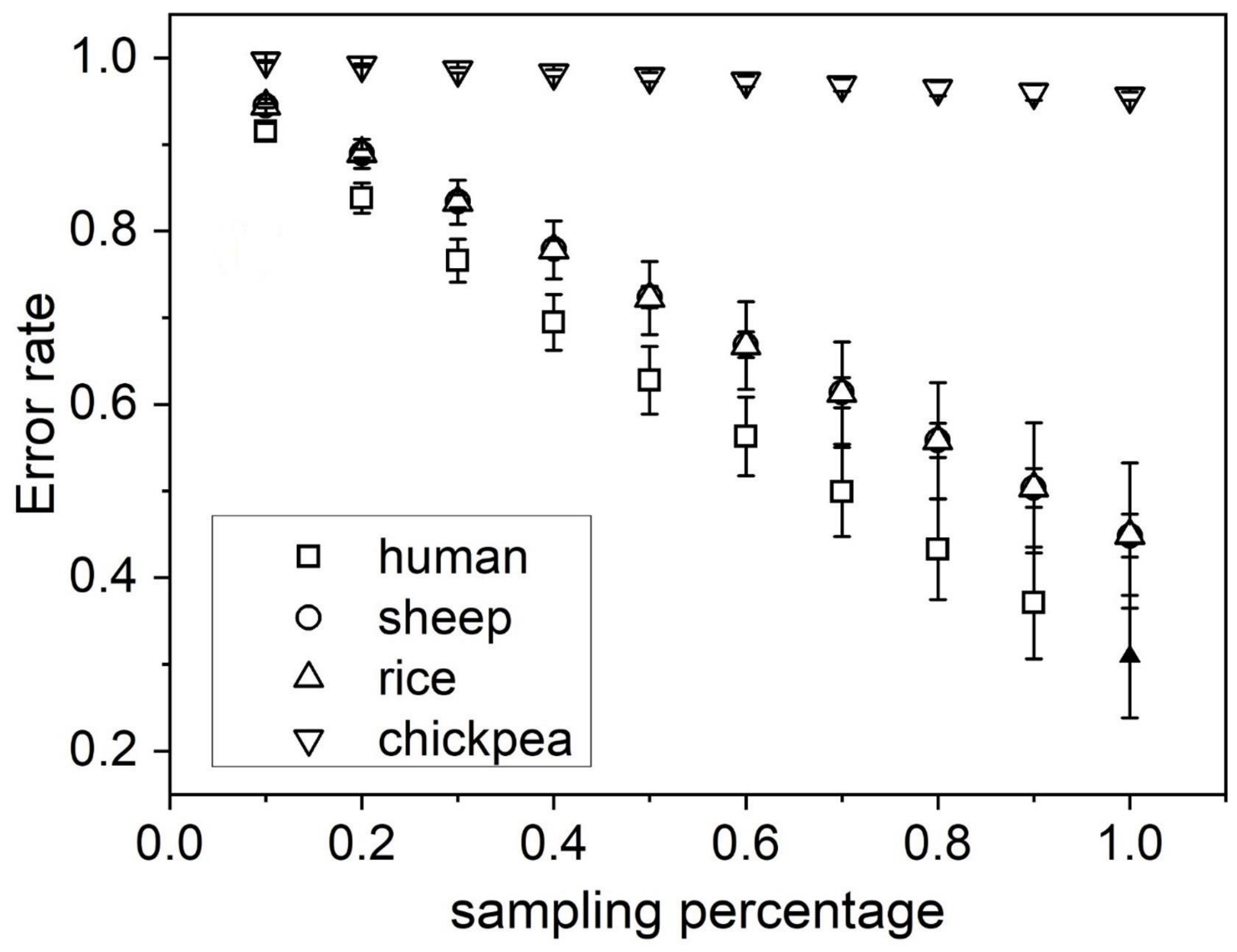

Fig 1. Error rates of four species versus different sample fractions.

- We investigated the effect of the database size in terms of the error rate.
- Error rate was defined as the ratio of the number of mismatched bases not listed in the dbSNP to the total number of mismatched bases in a sample.
- We constructed 10 test databases of different sizes, each of which was composed of variants randomly selected from the dbSNP from 10% to 100% at 10% interval.
- For all species, the error rate decreased as the ratio increases and there existed a gap between the error rate of human and those of other species [Fig. 1].
- This is because the number of variants in the dbSNP of other species is either not large enough or smaller than that of the human dbSNP.
- These results suggests that we need to construct 'pseudo database' for the species other than human.

3. Construction of pseudo-database

- We suggest a method of constructing a database when there is no database or the existing database does not contain enough number of variants [Fig. 2].

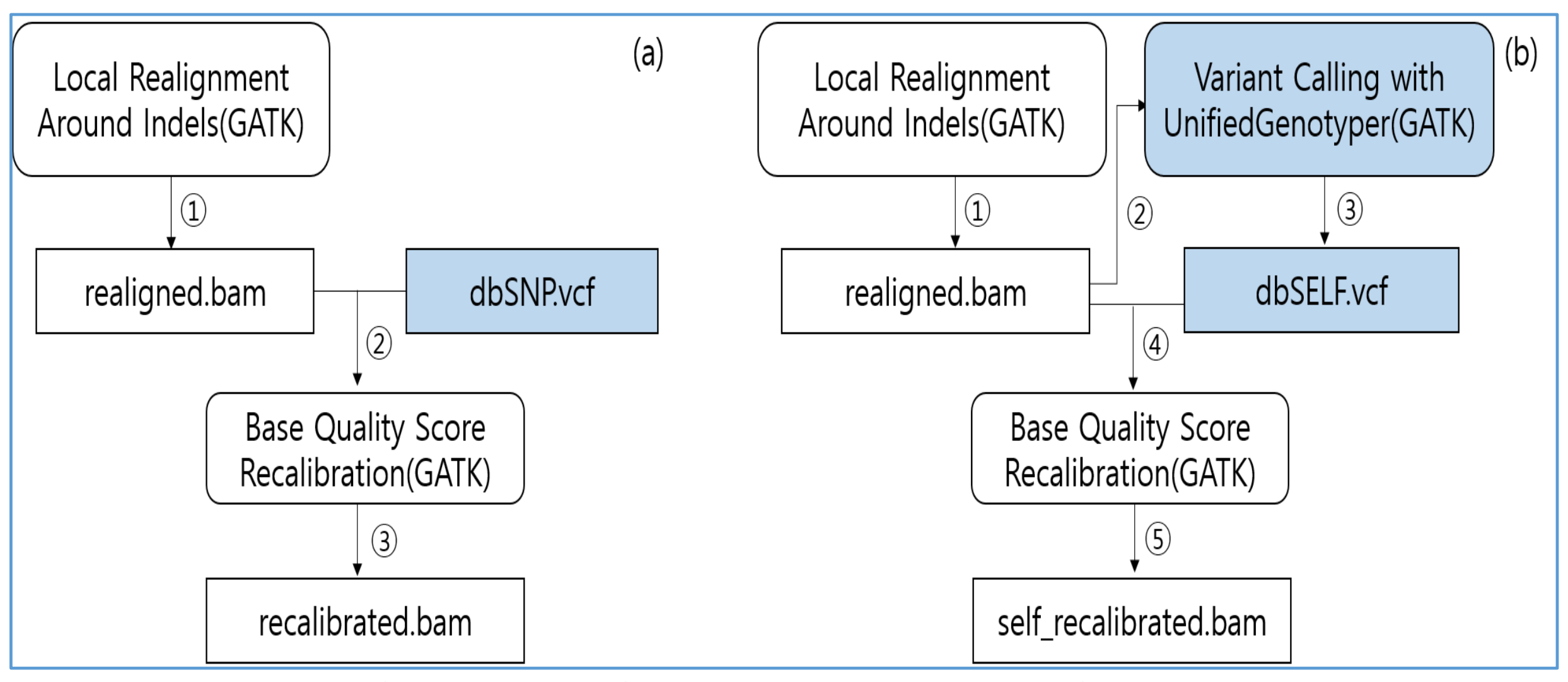
(a)
Local Realignment Around Indels(GATK)
①
realigned.bam
dbSNP.vcf
②
Base Quality Score Recalibration(GATK)
③
recalibrated.bam
(b)
Local Realignment Around Indels(GATK)
Variant Calling with UnifiedGenotyper(GATK)
①
②
③
realigned.bam
dbSELF.vcf
④
Base Quality Score Recalibration(GATK)
⑤
self_recalibrated.bam

Fig 2 : The schematic flow chart of current BQSR step(left) and the proposed step(right).

- Step 1: Perform variant calling by using a variant calling pipeline, such as GATK, without BQSR step to obtain a variant call format file(VCF) that contains variants called by the pipeline without BQSR step.
- Step 2: Perform the variant calling again including BQSR step by using the VCF as the database for the known variants.

III. Experimental result

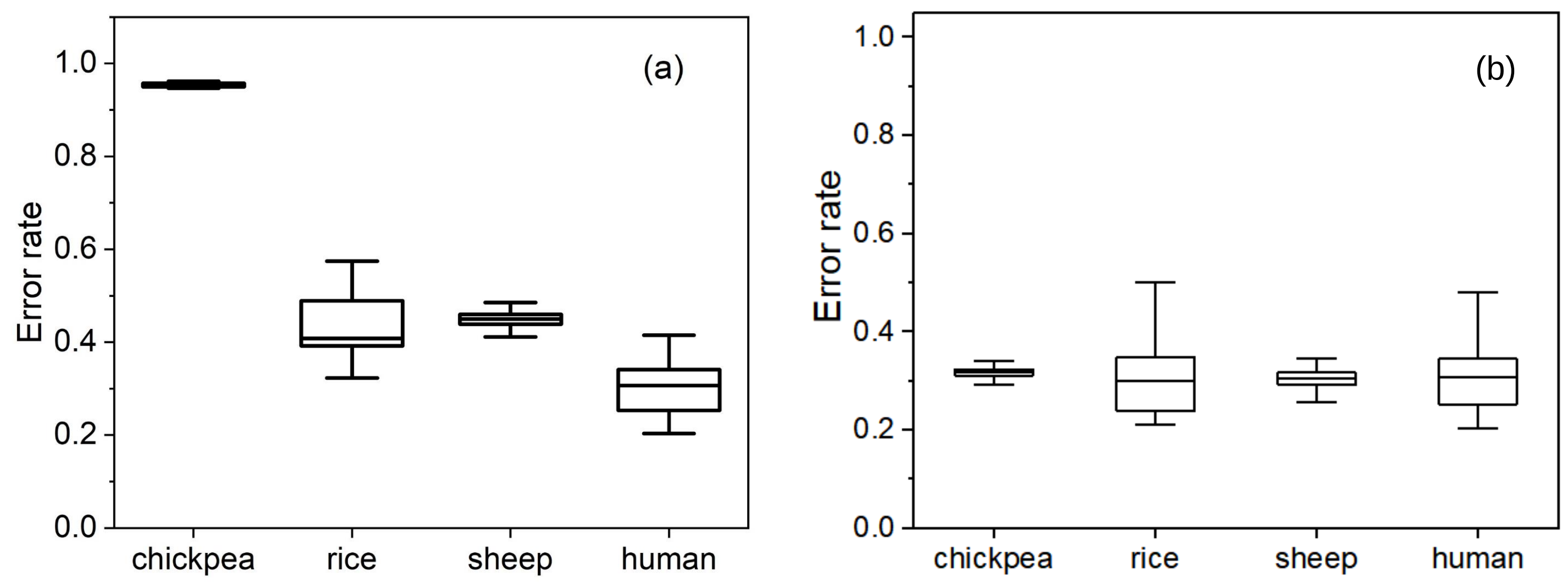

Fig 3 : (a) Box plots of the error rate estimated by using the latest build dbSNP of four species. (b) Box plots of error rates of chickpea, rice and sheep with pseudo database constructed by using 150, 15 and 10 samples respectively, together with that of human with dbSNP.

- Error rate estimated by using the pseudo database is comparable with that by using the human dbSNP [Fig. 3].
- While the pseudo database of sheep (or rice) required 10 (or 15) samples, that of chickpea required 150 samples.
- This is because chickpea reference sequence covers about 47% of its genome, while those of sheep and rice cover 94% and 98% of their genomes.

IV. Conclusion

- The recalibration results were closely relative to the size of dbSNP.
- We proposed a method to create a pseudo database when the size of the dbSNP is not large enough.
- In the case of sheep, rice and chickpea, the proposed method of construction the pseudo database is more reasonable than their dbSNP in the variant calling.
- The proposed method can be applied to the variant callings of other species for which the size of the database is not large enough.

V. References

- Brockman, W. and et al. (2008). Quality scores and snp detection in sequencing-by-synthesis systems. *Genome Res.*, 18(5):763–770.
- Cabanski, C. and et al. (2012). Reqon: a Bioconductor package for recalibrating quality scores from nextgeneration sequencing data. *BMC Bioinformatics*, 13:221.
- Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res.*, 8(3):186–194.

(*) This research was supported by (NRF-2021R1I1A3044289)